finding_john_smith

February 5, 2018

```
In [1]: import pandas as pd
        import numpy as np
        import re,os
        from sklearn import preprocessing
        %matplotlib inline
```

```
In [2]: %%time
        data = pd.read_csv('owners.csv')
```

```
CPU times: user 894 ms, sys: 135 ms, total: 1.03 s
Wall time: 1.08 s
```

```
In [3]: owner=list(data['OWNER'].astype('str'))
```

```
In [4]: count = 0
        names = []
        for name in owner:
            if 'john' in name.lower() and 'smith' in name.lower() :
                count+=1
                names.append(name)
        print(count)
```

```
43
```

```
In [5]: #this mehod includes johnson and john smith, as long as something else like HYSMITH, JOH
        names
```

```
Out[5]: ['JOHN SMITH',
         'SMITH, JOHN H',
         'JOHN SMITH',
         'SMITH, JOHN L',
         'JOHN L SMITH',
         'SMITH, JOHNNY',
         'SMITH, JOHN',
         'SMITH, JR., JOHN L',
         'SMITH JOHN J',
         'JOHN J SMITH',
         'JOHN SMITH',
         'JOHN L SMITH',
         'SMITH, JOHN CARL',
         'JOHN W SMITH',
         'SMITH JOHN',
         'SMITH, JOHN P',
         'SMITH, JOHN E',
         'JOHN SMITH JR',
         'SMITH JOHN R',
         'SMITH JOHN',
         'SMITH, JOHN W',
         'JOHN SMITH',
         'SMITH, JOHN',
         'JOHN A SMITH',
         'SMITH, GREOGROY JOHN',
         'SMITH, JOHN L',
         'JOHN SMITH',
         'JOHN L SMITH',
         'JOHN C SMITH',
         'JOHN L SMITH',
         'JOHN T SMITH',
         'SMITH JOHNSON, CAROL',
         'JOHN P SMITH',
         'JOHN SMITH',
         'JOHN SMITH',
         'SMITH JOHN',
         'JOHN L SMITH',
         'HYSMITH, JOHN W',
         'JOHN W SMITH',
         'JOHN W SMITH',
         'SMITH, JOHN H',
         'SMITH, JOHN OLIVER',
         'SMITH, JOHN, C.']
```

```
In [6]: from difflib import SequenceMatcher
        def similar(a, b):
            return SequenceMatcher(None, a, b).ratio()
```

```
In [15]: #this process is quite slow though
         count_matcher = 0
         names_matcher = []

         for name in owner:
             name_cleaned = name.lower().replace(',','').replace(' ','')
             if any([similar(name_cleaned, 'johnsmith') > 0.8, similar(name_cleaned, 'smithjohn'
                    similar(name_cleaned, 'johnnysmith') > 0.8, similar(name_cleaned, 'smithjohn
                 count_matcher+=1
```

```
                names_matcher.append(name)
        print(count_matcher)
```

```
51
```

```
In [16]: names_matcher
```

```
Out[16]: ['JOHN SMITH',
          'SMITH, JOHN H',
          'JOHN SMITH',
          'SMITH, JOHN L',
          'JOHN L SMITH',
          'SMITH, JOHNNY',
          'SMITH, JOHN',
          'SMITH, JR., JOHN L',
          'SMITH JOHN J',
          'JOHN J SMITH',
          'SMITH, JOAN E',
          'JOHN SMITH',
          'JOHN L SMITH',
          'SMITH, JOAN',
          'SMITH, JOAN T',
          'SMITH, SONNY',
          'SMITH, JOE',
          'SMITH, JOHN CARL',
          'JOHN W SMITH',
          'SMITH JOHN',
          'SMITH, JOHN P',
          'SMITH, JOHN E',
          'SMITH, JOAN P',
          'JOHN SMITH JR',
          'SMITH JOHN R',
          'SMITH JOHN',
          'SMITH, JOHN W',
          'JOAN SMITH',
          'JOHN SMITH',
          'SMITH, JOHN',
          'JOHN A SMITH',
          'SMITH, JOHN L',
          'JOHN SMITH',
          'JOHN L SMITH',
          'JOANNE SMITH',
          'JOHN C SMITH',
          'RON SMITH',
          'JOHN L SMITH',
          'JOHN SAITH',
          'JOHN T SMITH',
          'JOHN P SMITH',
          'JOHN SMITH',
          'SMITH JOANNE',
          'JOHN SMITH',
          'SMITH JOHN',
          'JOHN L SMITH',
          'HYSMITH, JOHN W',
          'JOHN W SMITH',
          'JOHN W SMITH',
          'SMITH, JOHN H',
          'SMITH, JOHN, C.']
```

```
In [17]: import nltk
```

```
In [18]: count_nltk = 0
         names_nltk = []

         #nltk takes time as well

         for name in owner:
             token = nltk.word_tokenize(name.lower())
             if ('john' in token and 'smith' in token) or ('johnny' in token and 'smith' in toke
                 count_nltk += 1
                 names_nltk.append(name)
         print(count_nltk)
```

```
41
```

```
In [19]: names_nltk
```

```
Out[19]: ['JOHN SMITH',
          'SMITH, JOHN H',
          'JOHN SMITH',
          'SMITH, JOHN L',
          'JOHN L SMITH',
          'SMITH, JOHNNY',
          'SMITH, JOHN',
          'SMITH, JR., JOHN L',
          'SMITH JOHN J',
          'JOHN J SMITH',
          'JOHN SMITH',
          'JOHN L SMITH',
          'SMITH, JOHN CARL',
          'JOHN W SMITH',
          'SMITH JOHN',
          'SMITH, JOHN P',
          'SMITH, JOHN E',
          'JOHN SMITH JR',
          'SMITH JOHN R',
          'SMITH JOHN',
          'SMITH, JOHN W',
          'JOHN SMITH',
          'SMITH, JOHN',
          'JOHN A SMITH',
          'SMITH, GREOGROY JOHN',
          'SMITH, JOHN L',
          'JOHN SMITH',
          'JOHN L SMITH',
          'JOHN C SMITH',
          'JOHN L SMITH',
          'JOHN T SMITH',
          'JOHN P SMITH',
          'JOHN SMITH',
          'JOHN SMITH',
          'SMITH JOHN',
          'JOHN L SMITH',
          'JOHN W SMITH',
          'JOHN W SMITH',
          'SMITH, JOHN H',
          'SMITH, JOHN OLIVER',
          'SMITH, JOHN, C.']
```